Topic Modeling the Human Plasma Proteome: An Unsupervised Learning Method for Proteomic Analysis

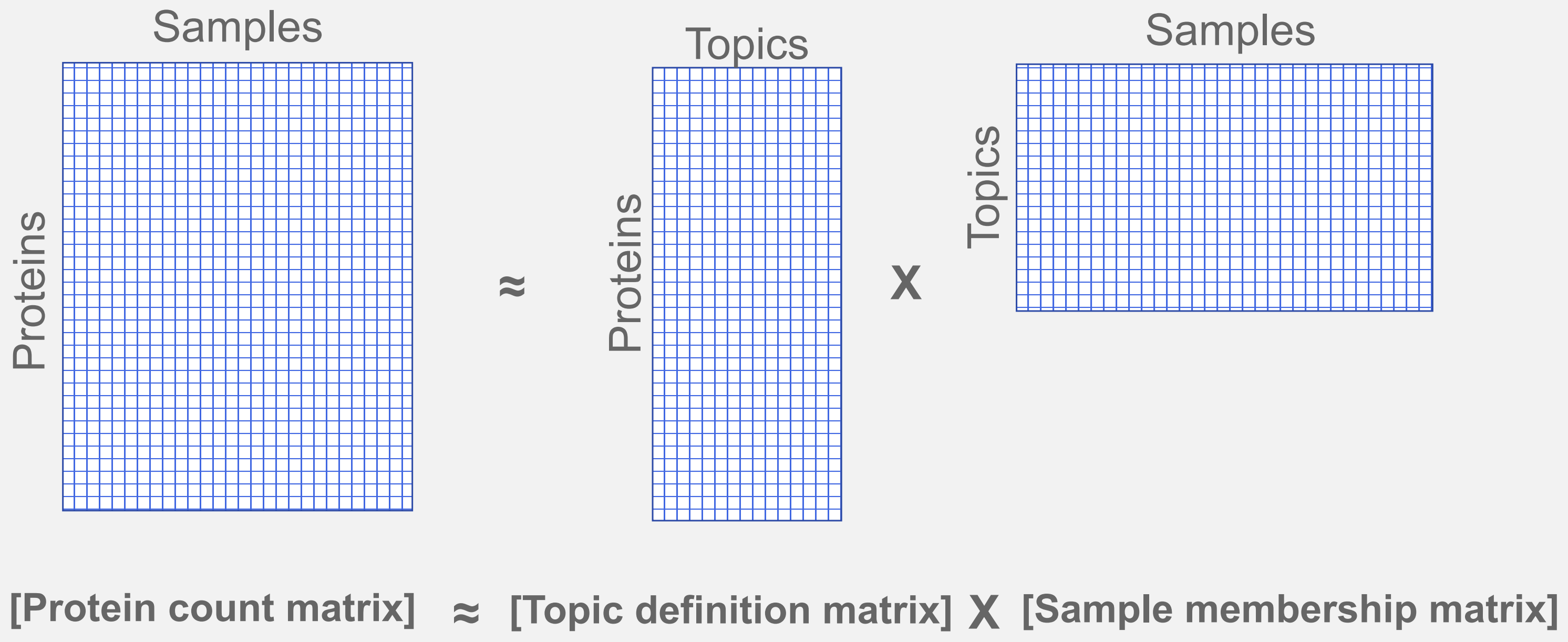
Blake J.M. Williams MS, R. Kirk DeLisle PhD

Background

Goal: Identify interpretable dimensions in 7k SomaScan Assay using **topic models**, identify disease subpopulations, investigate relationships between topics and clinical measurements

Methods: Apply topic models using non-negative matrix factorization (NMF), on non-alcoholic fatty liver disease (NAFLD) samples

Methods

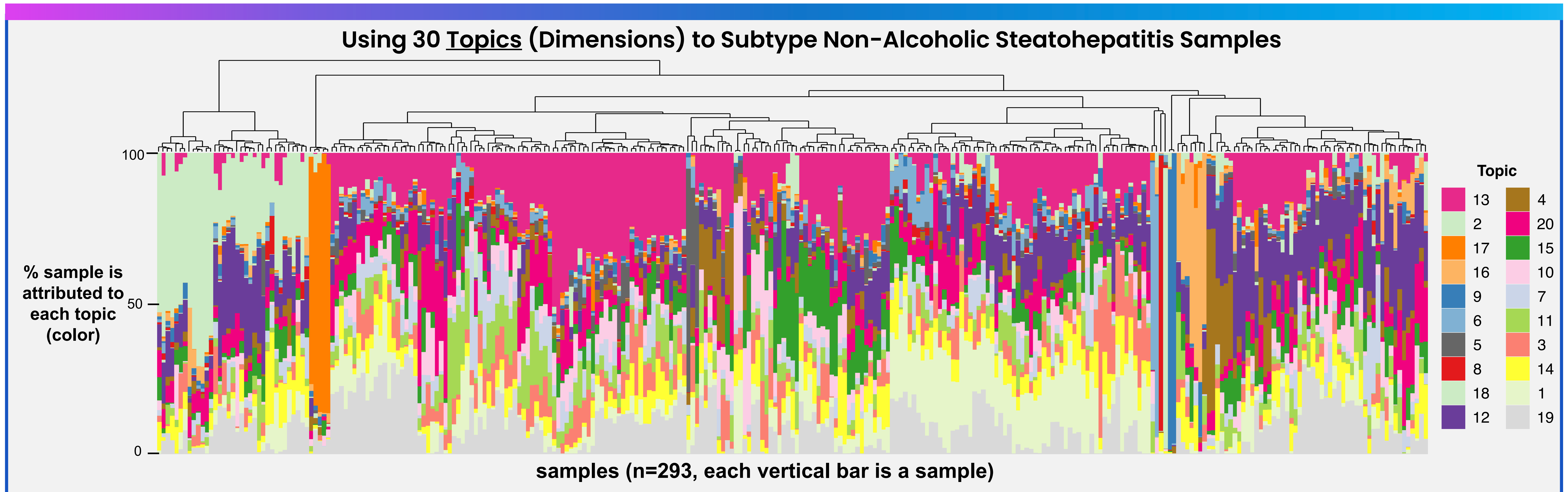


[Protein count matrix] ≈ [Topic definition matrix] X [Sample membership matrix]

- **Decompose [sample x topic] matrix into two matrices**, whose product best reproduces original matrix:
  - Fit using NMF
- **Model output matrices:**
  - **[Protein x topic]**: defines topics w/protein weights
  - **[Topic x sample]:** attributes some % of sample to each topic
- **Advantages:**
  - **Sparsity**: ~50-200 proteins describe each topic (dimension), assists in interpretability of topic biological meaning
  - **Non-orthogonal dimensions:** better distinguish similar/ overlapping functions

Results

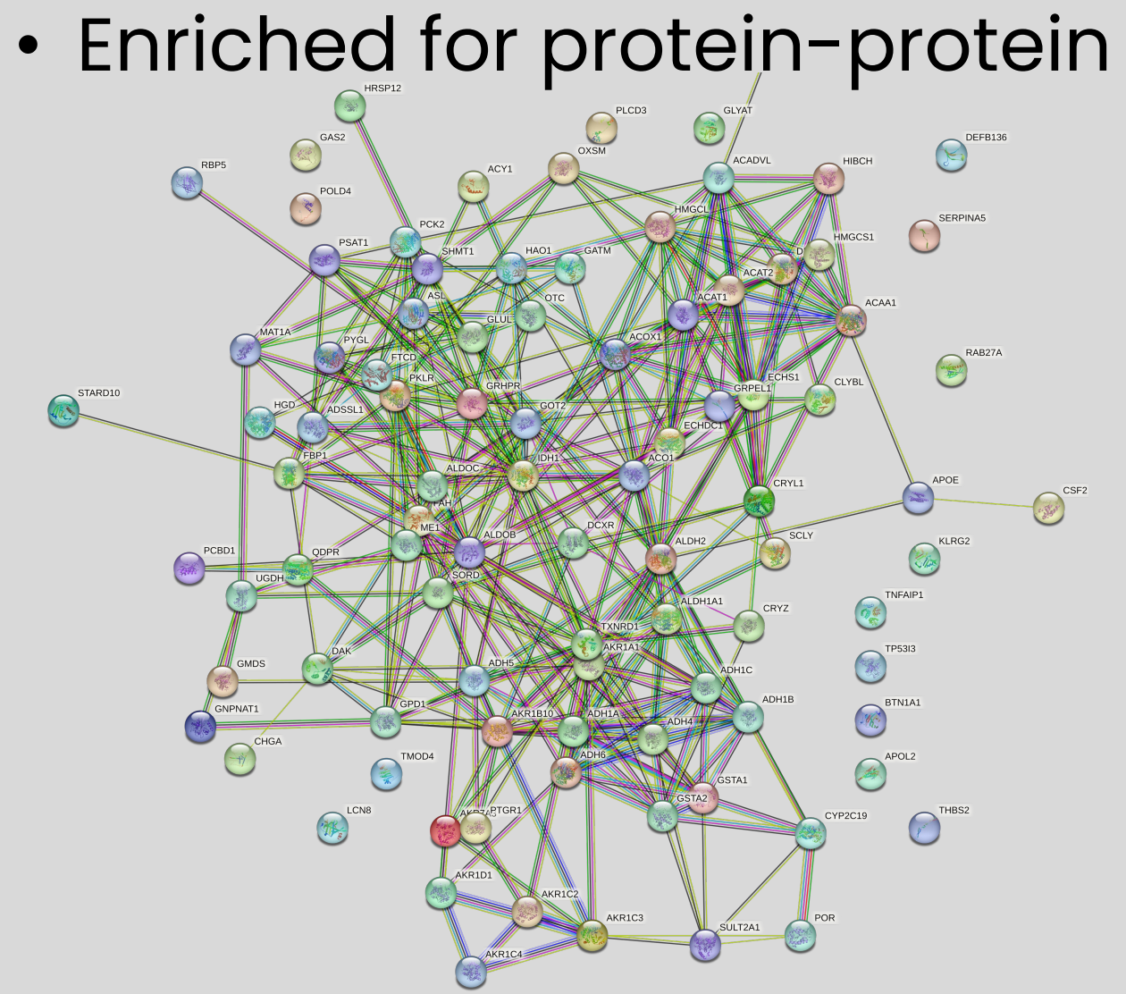
Using 30 Topics (Dimensions) to Subtype Non-Alcoholic Steatohepatitis Samples



samples (n=293, each vertical bar is a sample)

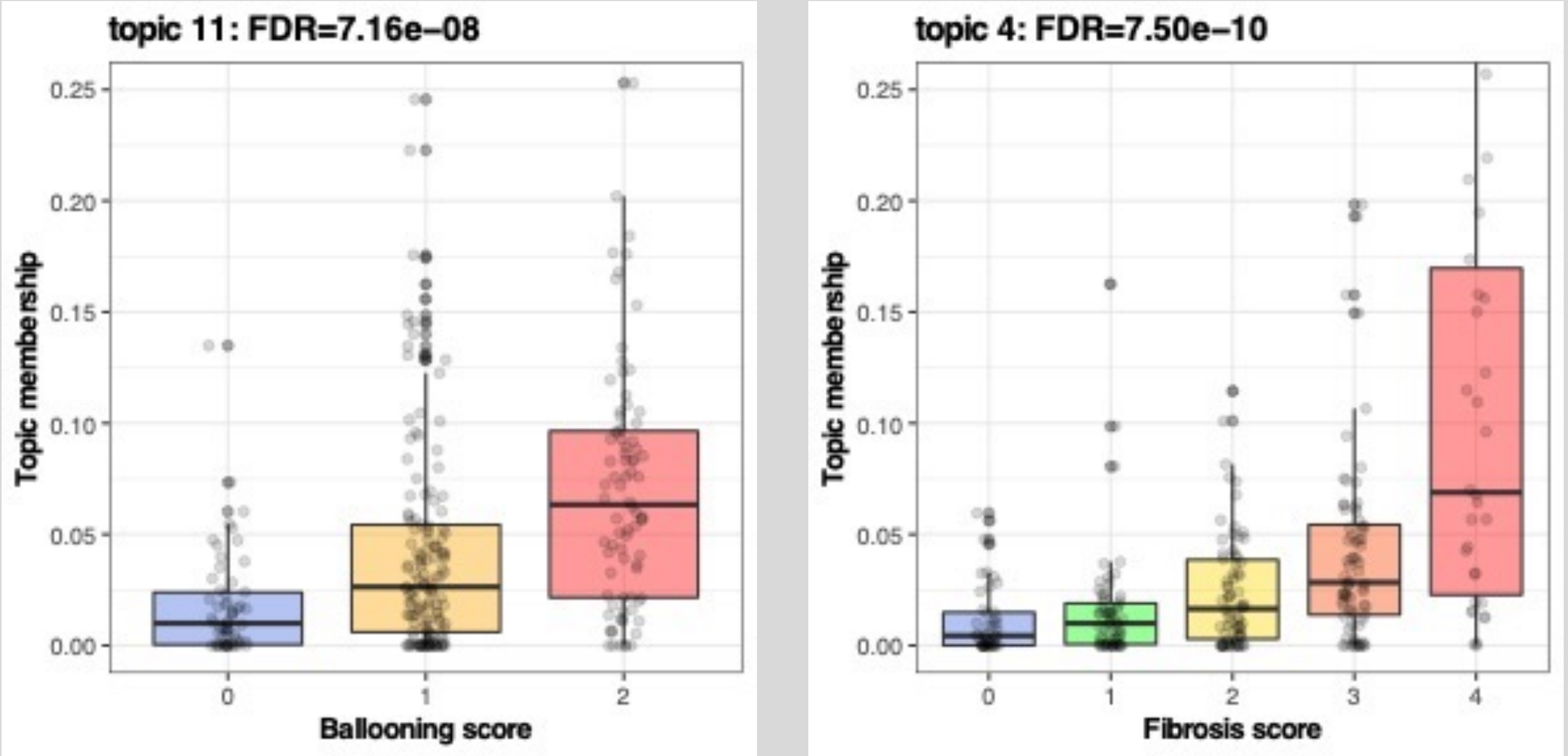
Topics (dimensions) are biologically coherent

- Topics capture proteomic pathways
  - E.g. topic 11:
    - Hepatic steatosis (FDR=9E-06, hypergeometric test)
    - Proteins expressed in liver (FDR-3E-40, hypergeometric test)
- Proteins are ranked in topics by their weights
- Top proteins in topics:
  - Include related protein biomarkers
  - Are pathway-enriched for pathways
  - Enriched for protein-protein interactions



Dense PPI network of fatty liver-related topic 15

Topics identify clinically relevant dimensions



- **Topics are correlated with clinical measurements:**
  - Topic 11: cellular ballooning (FDR=7.2E-08, hypergeometric test)
  - Topic 4: fibrosis score (FDR=7.5E-10, hypergeometric test)
- These topics suggest a biological explanation for correlated traits

Contact Us

Scan the QR code for more information!

Conclusions

- **Clinically relevant & interpretable dimensions** are identified by topic models of SomaScan assay
- **Disease subgroups** are identified by clustering on topic membership
- **Biological explanation for disease subtypes** suggested by pathway analysis of correlated topics

Acknowledgements

Acknowledgement

The LITMUS project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 777377. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

Disclaimer

This communication reflects the view of the author(s) and neither IMI nor the European Union or EFPIA are liable for any use that may be made of the information contained herein.

www.litmus-project.eu
www.imi.europa.eu

© 2022 SomaLogic, Inc. All rights reserved.